SECOND MEETUP

DiSCo

BGU's Data Science Community

MEET • TEAM UP • KAGGLE

HAVE FUN!

DiSCo
Data Science Community

Bengis Center for Entrepreneurship & Innovation
Guilford Glazer Faculty of Business and Management
Ben-Gurion University of the Negev

BEN-GURION UNIVERSITY OF THE NEGEV

8 APRIL 2018

DOORS OPEN AT 18:30

> Data Preprocessing
> Feature transformation and feature engineering
> Improving performance of regression model by feature selection

kaggle Search kaggle Competitions Datasets Kernels Discussion Jobs

House Prices: Advanced Regression Techniques

Predict sales prices and practice feature engineering, RFs, and gradient boosting

4,375 teams · 2 years to go

Overview Data Kernels Discussion Leaderboard Rules Team My Submissions Submit Predictions

- Description
- Evaluation
- Frequently Asked Questions
- Tutorials

Start here if...

You have some experience with R or Python and machine learning basics. This is a perfect competition for data science students who have completed an online course in machine learning and are looking to expand their skill set before trying a featured competition.

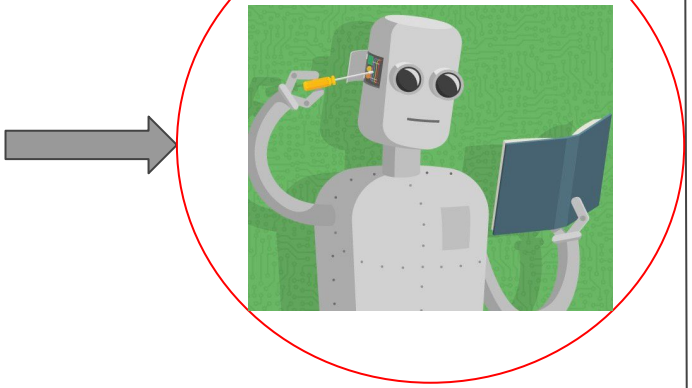
Competition Description

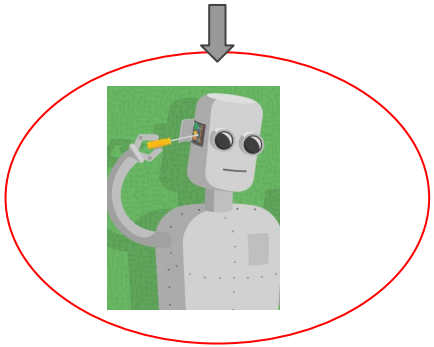
Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.

With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition challenges you to predict the final price of each home.

House Prices Competition

Problem Statement



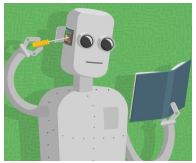


Features of individual houses

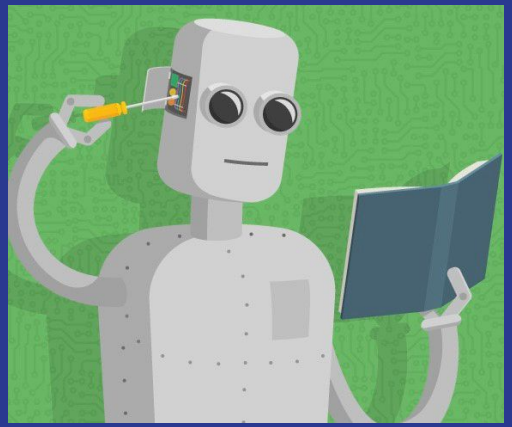
What factors can you think of right now which can influence house prices ?

- Area of House
- How old is the house
- Location of the house
- If terrace is available
- If car parking is available
- If security is available

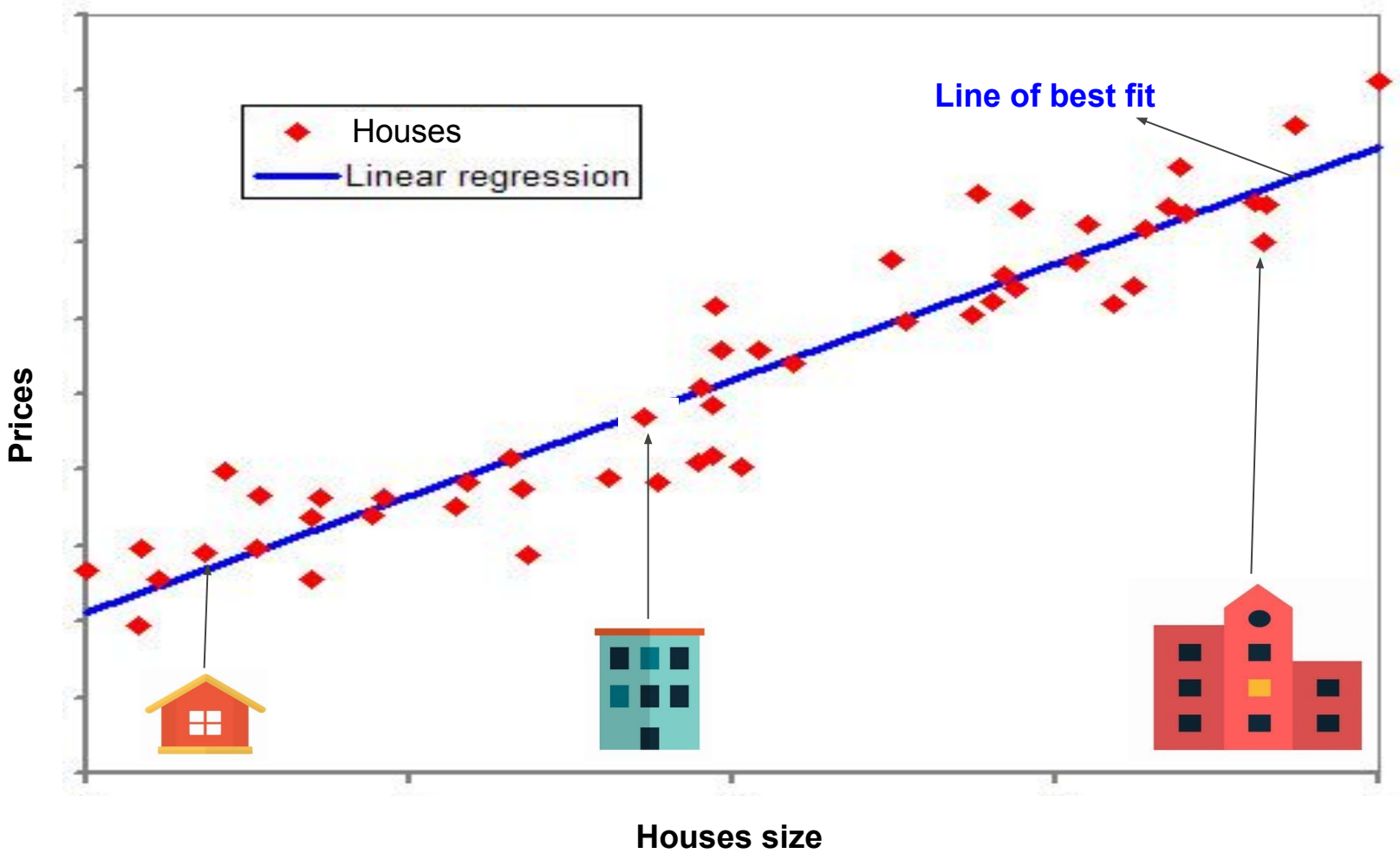
- Price



Simple Linear regression



Express-highway of best fit


Houses
Linear regression
Line of best fit
Prices
Houses size

Linear regression : Introduction

In supervised learning, we have examples of lots of input and the desired output value.

This is called the **dataset**:

- A matrix of **feature vectors** X.
- A vector of **target values** Y.

- Area of House
- How old is the house
- Location of the house
- If terrace is available
- If car parking is available
- If security is available

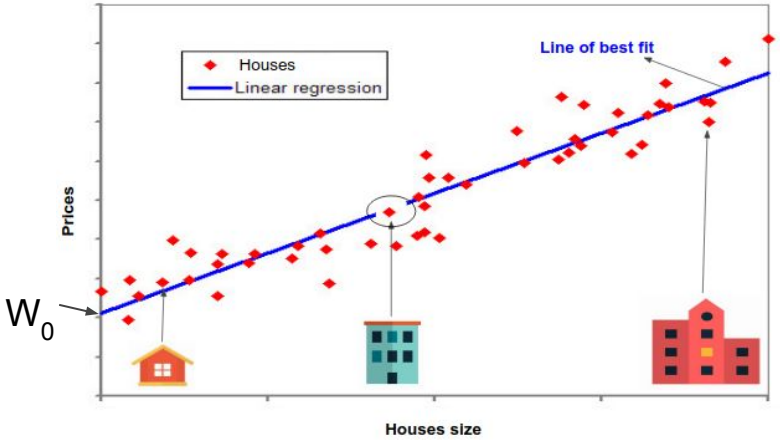
- Price

Linear model

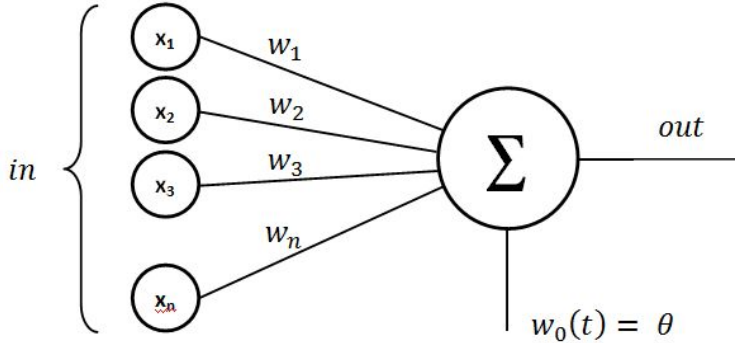
Coefficient

$$Y = w_0 + w_1 * x1$$

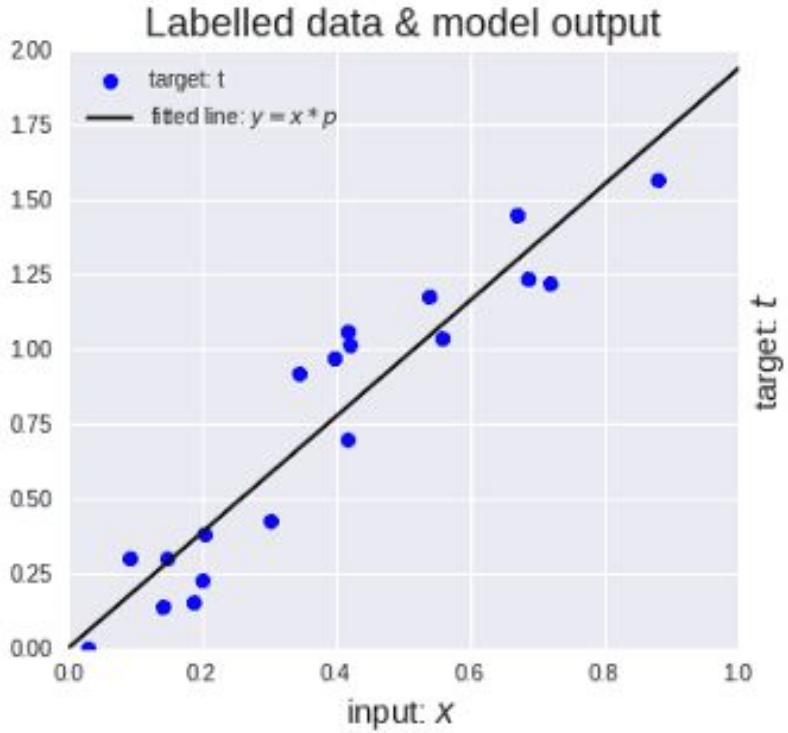
Dependent variable (DV) Independent variable (IV)



W_1 = It is the change in y for a unit change in x along the line.



3d visualization of multiple regression

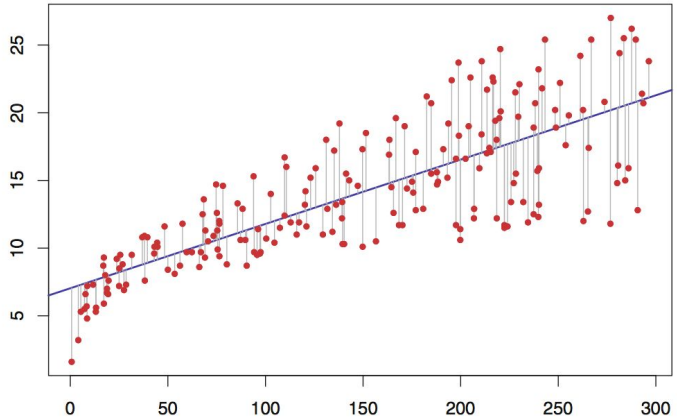


Linear regression : Mean Squared Error

Euclidean distance between the target and predicted.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

$$E(W) = \frac{1}{2}\sum_{i=0}^{n}(W.X_i^T - Y_i)^2$$



MSE(Mean Squared Error) and Error Function

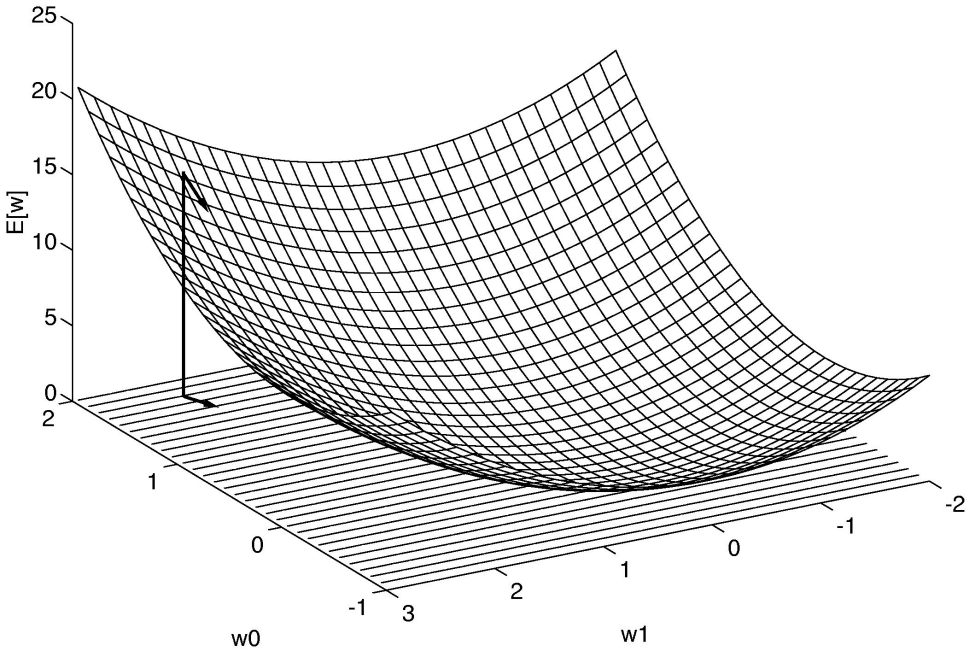
Linear regression : gradient descent

$$E(W) = \frac{1}{2}\sum_{i=0}^{n}(W.X_i^T - Y_i)^2$$

The least squared loss of a linear model is a convex function (“bowl-shaped”)

One simple way to find its minimum is by **following the slope of the error**.

$$W \leftarrow W + \alpha\frac{\delta E}{\delta W}$$



Gradient Descent by Andrew Ng:
https://www.youtube.com/watch?v=yFPLyDwVifc

Batch and stochastic gradient descent : https://tinyurl.com/y8vrt6qo

Illustration: Tom Mitchell, McGraw-Hill

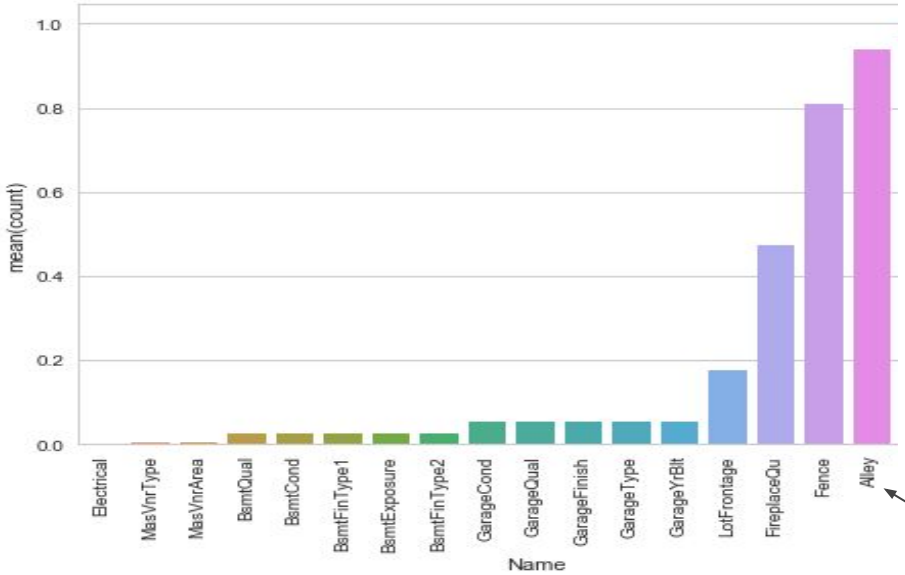
Data Exploration

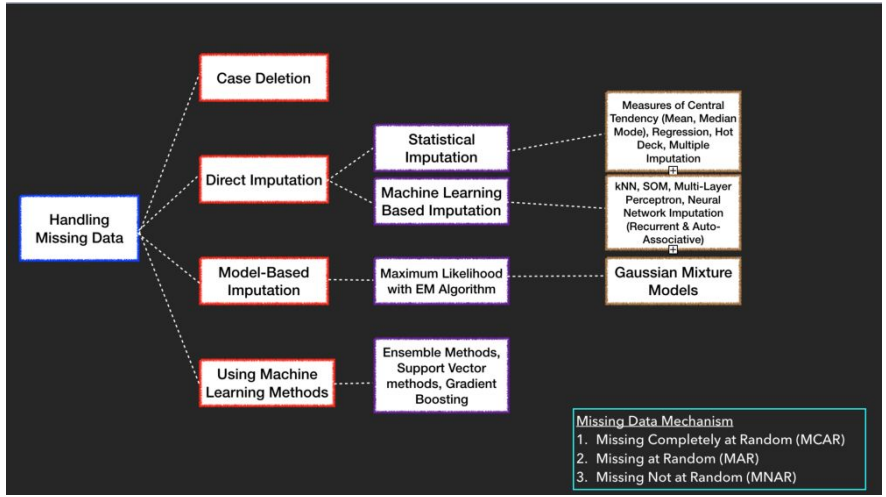
Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities
1	60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPub
2	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub
3	60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPub
4	70	RL	60.0	9550	Pave	NaN	IR1	Lvl	AllPub
5	60	RL	84.0	14260	Pave	NaN	IR1	Lvl	AllPub

Statistical Analysis

1. **Univariate Analysis** - Examples: histogram, density plot, etc.
2. **Bivariate Analysis** - Examples: bar chart, line chart, area chart, etc.
3. **Multivariate Analysis** - Examples: stacked bar chart, dodged bar chart, etc.

Univariate Analysis - Plot the missing values count for features

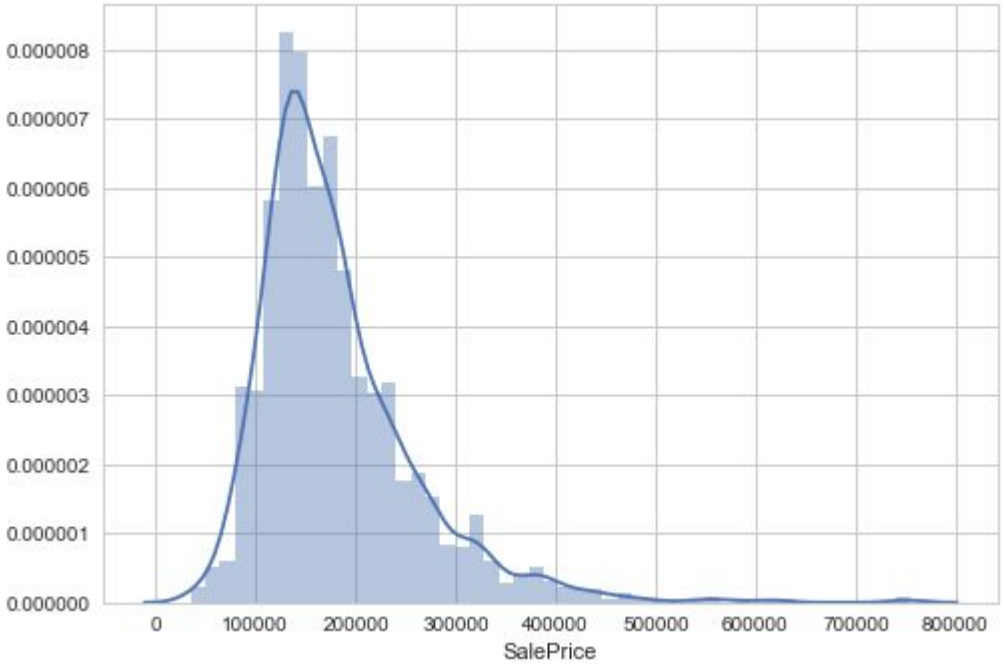




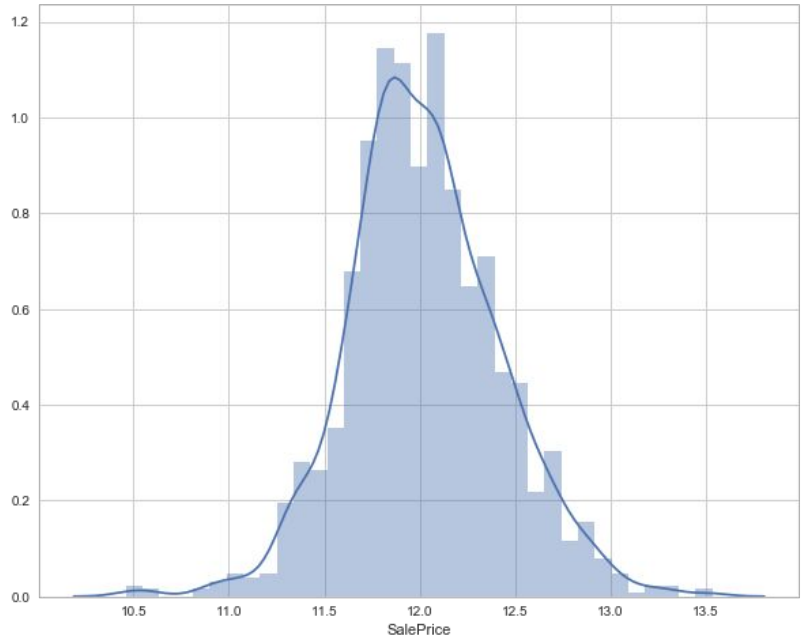
Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities
1	60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPub
2	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub
3	60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPub
4	70	RL	60.0	9550	Pave	NaN	IR1	Lvl	AllPub
5	60	RL	84.0	14260	Pave	NaN	IR1	Lvl	AllPub

Univariate Analysis - Transformation of 'Target Variable' - Sales Prices

- Skewness : A measure of assymetry in the distribution
- SalesPrices distribution is concentrated to the left

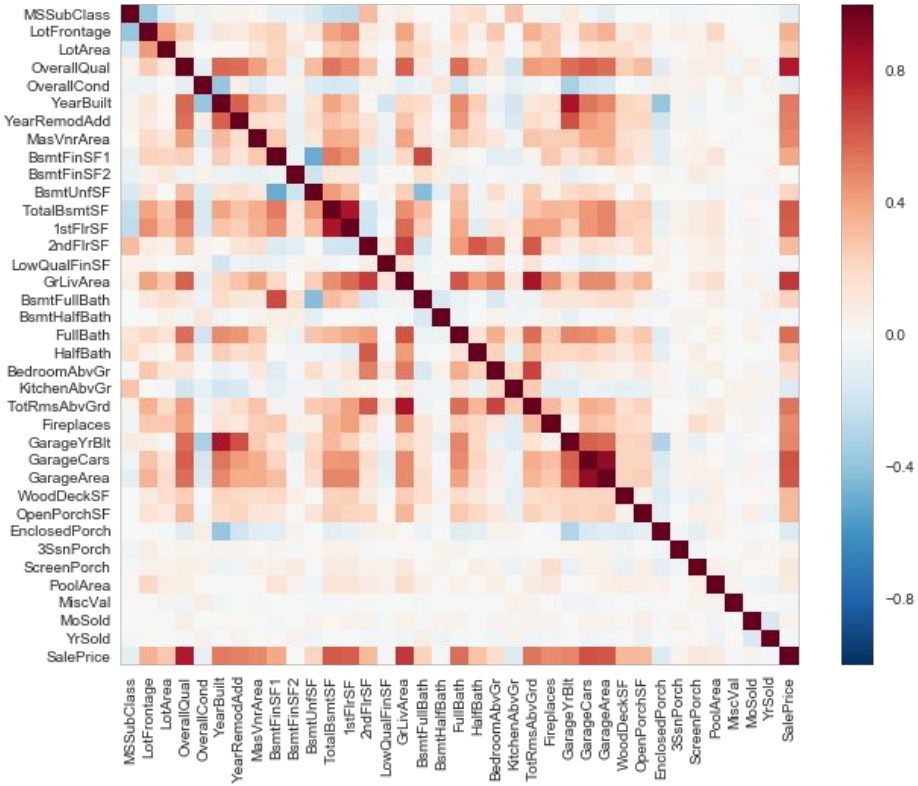


Positively Skewed distribution

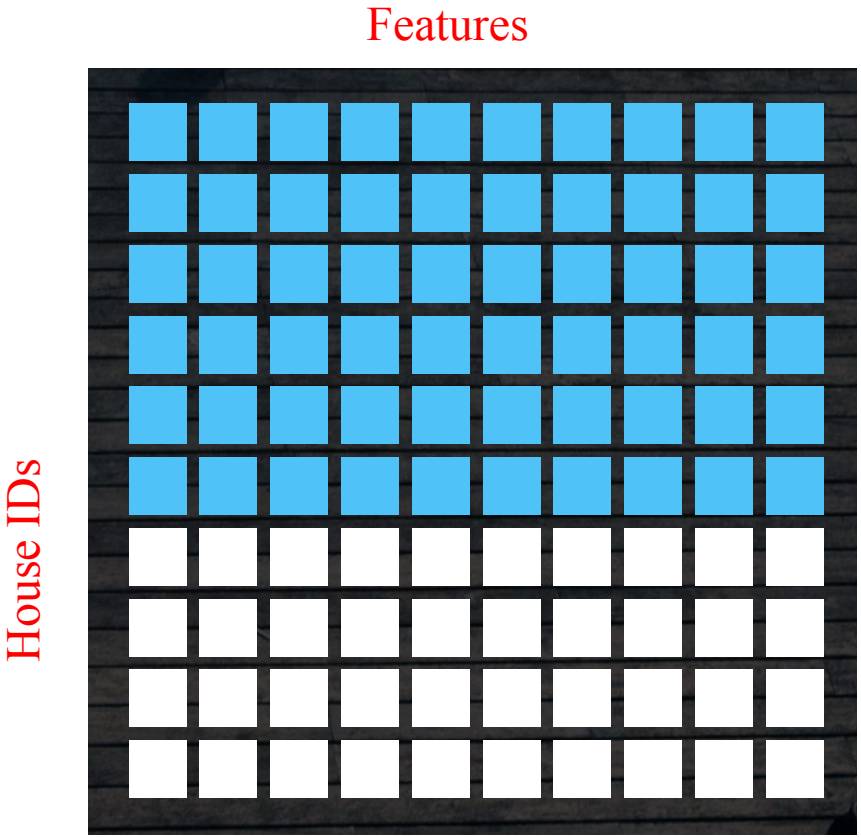


Normal Distribution

Bivariate Analysis - Correlation heatmap of Salesprice with all other features



ML workflow



Training Set

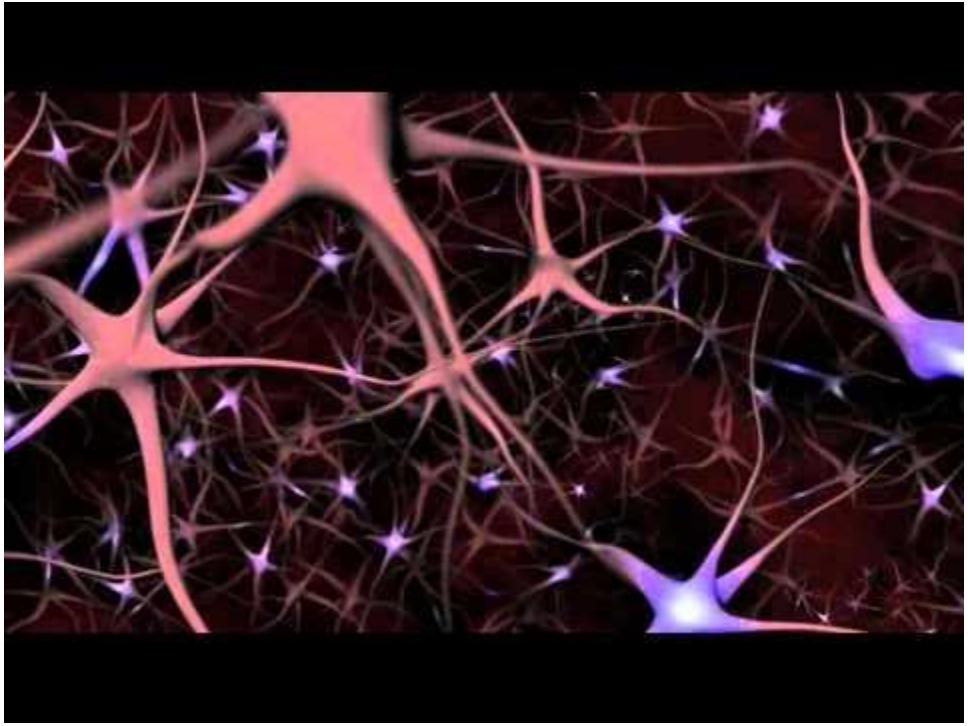
Test Set

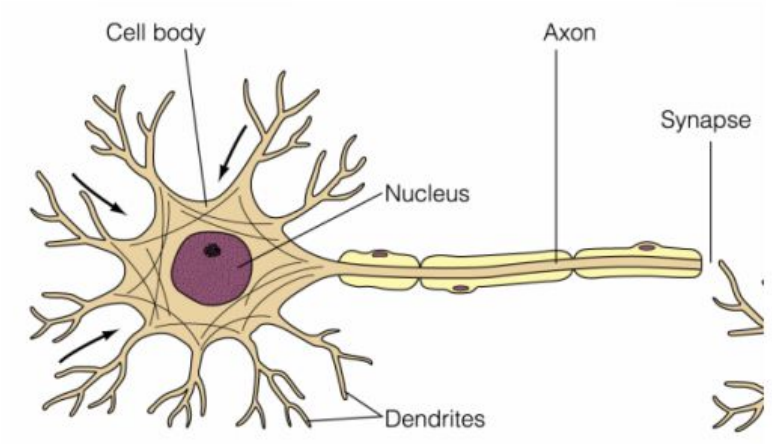
Model Building using training set

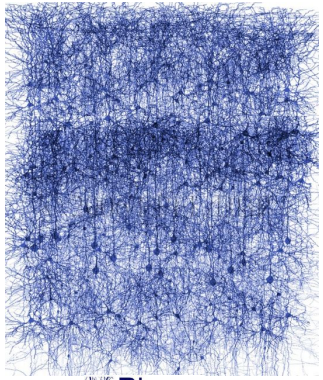
Cross Validation

Prediction using test set

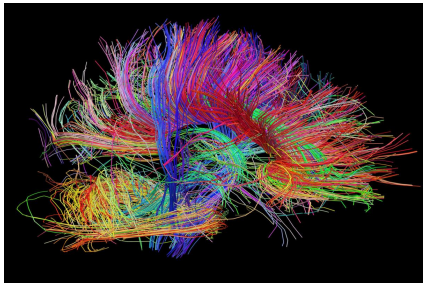
Introduction to neural networks



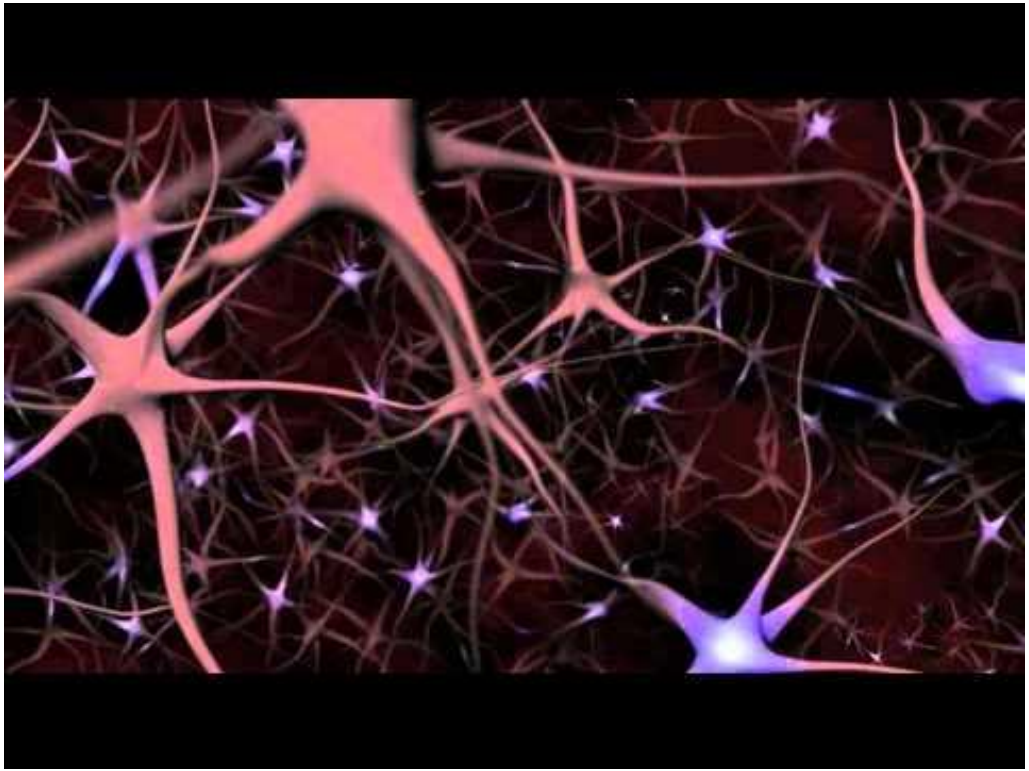


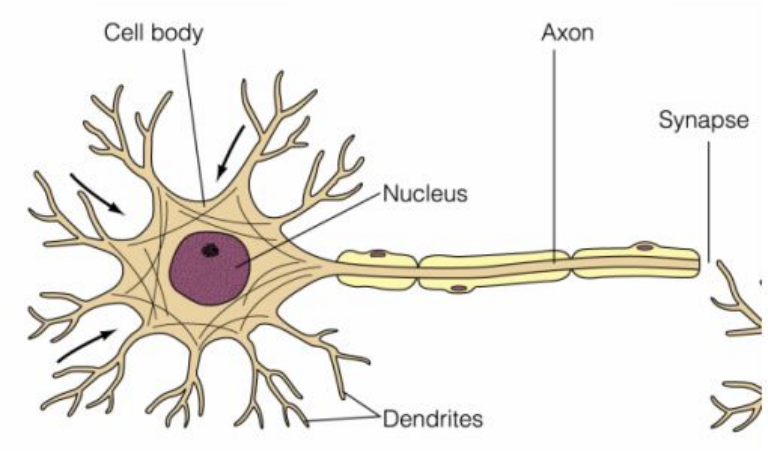


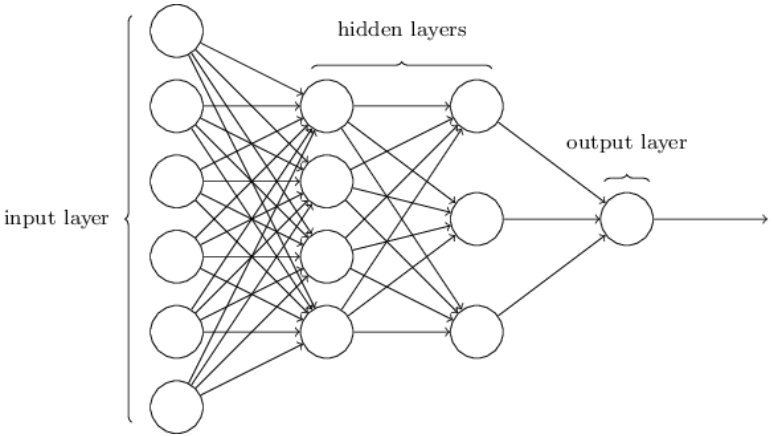
Blue Brain Project



Introduction to neural networks





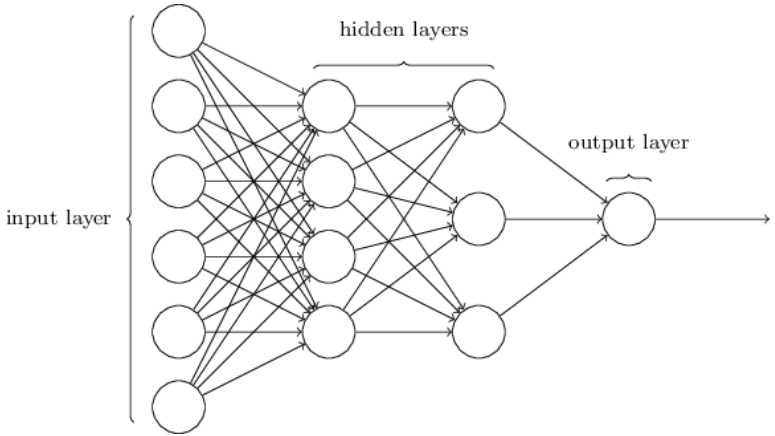


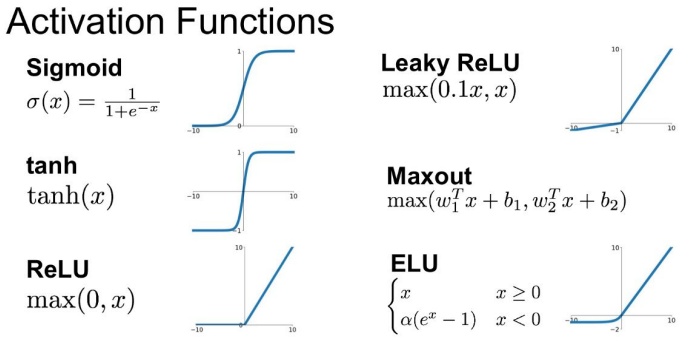
Introduction to activation functions

A function that helps neuron to decide whether to fire or not

Or

A function used to transform the activation level of neuron (weighted sum of inputs) to an output signal.



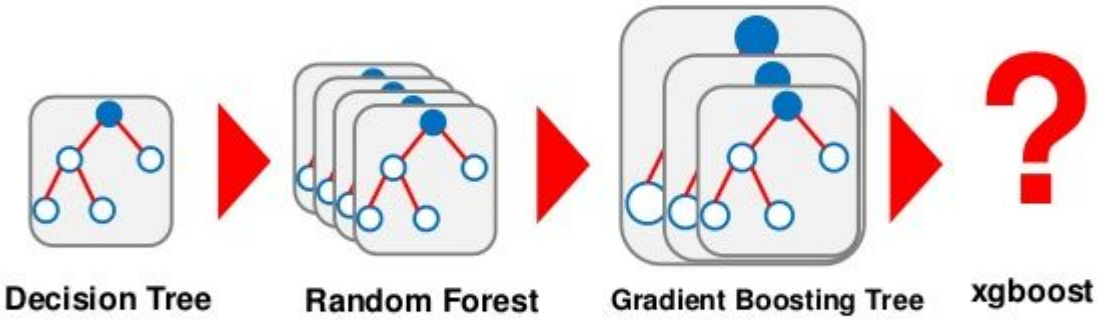


Ref:
http://neuralnetworksanddeeplearning.com/chap1.html

More on Activation function: https://tinyurl.com/ybrfnh24

XGBoost (extreme gradient boosting)

XGBoost is a for Gradient boosting trees model



What's happened during this evolution?

XGBoost (extreme gradient boosting)

XGBoost is an implementation of gradient boosted decision trees designed for speed and performance.

XGBoost library implements the gradient boosting decision tree algorithm.

Features: Three main forms of gradient boosting are supported:

1. **Gradient Boosting** algorithm also called gradient boosting machine including the learning rate.
2. **Stochastic Gradient Boosting** with sub-sampling at the row, column and column per split levels.
3. **Regularized Gradient Boosting** with both L1 and L2 regularization.

Ref: https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/
Trevor Hastie https://youtu.be/wPqtzj5VZus
https://www.quora.com/What-is-the-difference-between-the-R-gbm-gradient-boosting-machine-and-xgboost-extreme-gradient-boosting

Ok! Let’s get our hands dirty

From Week 1

https://tinyurl.com/disco-git
https://tinyurl.com/disco-kernel1
https://tinyurl.com/disco-facebook

For Week 2

https://tinyurl.com/kernal-week2
https://tinyurl.com/week2-resources

Acknowledgement

Course Creation and people responsible

Github page, kernal for week 2 - Minesh Jethva

Administration - Moran Sharon

Statistical help - Ruth Hashkes

Week 2 presentation - Rahul Veettil

Team DiSCo

Rahul Veettil

Minesh Jethva

Ruth Hashkes

Moran Sharon

https://www.bengis.org/disco
https://discobgu.github.io/Preprocessing/
Contact : disco.bgu@gmail.com


Thank you!